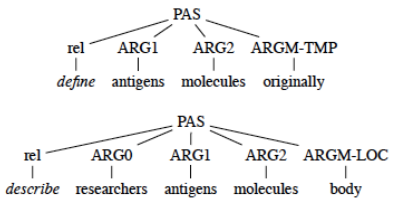

Figure 14. Shallow semantic tree

Query	Results	Normalized
President is a person	259	0.8662
President is a place	9	0.0301
President is an organization	11	0.0368
President is a measure	20	0.0669
President is a date	0	0

Figure 15. Example of using the Internet to extract features for question classification

information, could prevent the sparseness of deep structural approaches (syntactic parse tree) and the weakness of BOW models. They applied Semantic Role Labeling (SRL) [10] to QA system. By using PropBank(PB) [11], the Penn English Treebank with the addition of semantic information, SRL tasks can be done accurately. The goal is to label syntactic nodes with specific argument labels that preserve the similarity of roles such as *the window* in *John broke the window* and *the window broke*.

Consider the PB annotation: [ARG1 Antigens] were [AM-TMP originally] [rel defined] [ARG2 as non-self molecules]. Such annotation can be used to design a shallow semantic representation that can be matched against other semantically similar sentences, e.g. [ARG0 Researchers] [rel describe] [ARG1 antigens] [ARG2 as foreign molecules] [ARGM-LOC in the body]. They are represented in Predicate-Argument Structures (PAS) as shown in Figure 14.

The tree kernel explained in Subsection 5.3 is then applied to the PAS to be used as input to SVM.

5.3.3. Language independent method

The aforementioned approaches have the disadvantage of being targeted to a particular language. Solorio et al. [12] presented a simple approach that exploits lexical features and the Internet to train a SVM classifier. The main feature of this method is that it can be applied to different languages without requiring major modifications.

The procedure for gathering the information from the web is as follows: a set of heuristics is used to extract from the question a word *w*, or set of words, that will complement the queries submitted for the search. Then a search engine is used, in this case Google, and queries are submitted using the word *w* in combination with all the possible semantic classes. For instance, for the question *"Who is the President of the French Republic?"* the word President is extracted using heuristics, and then 5 queries are run in the search engine, one for each possible class. These queries take the following form:

- President is a person
- President is a place
- President is a date
- President is a measure
- President is an organization

The number of results returned by Google for each query is counted and normalized, as displayed in Figure 15. The resultant numbers are the values for the attributes used by the learning algorithm.

	H	C	H+C	U+H	U+C	U+H+C
Coarse	63.2%	92.0%	94.6%	91.8%	95.0%	95.0%
Fine	39.0%	83.4%	88.8%	84.2%	90.2%	90.8%

Figure 16. Accuracy of the hybrid classifier for coarse- and fine-grained categories

Approach	Coarse	Fine
Handwritten rules	87.0%	83.2%
Bag-of-words	85.8%	80.2%
Bag-of-ngrams	87.4%	79.2%
Syntactic parse tree	90.0%	80.2%
Hybrid	95.0%	90.8%

Figure 17. Accuracy of each classifier for coarse- and fine-grained categories

6. HYBRID APPROACH

Silva et al. [5] also proposed the hybrid approach. The information provided by the rule-based classifier - both headwords (H) and categories (C) - is used to generate the feature set for training, training, and merged with the information provided by the question unigrams (U).

It is when these features are combined with unigrams that the classifier holds the best results: an increase of 8.0 $¥%$ and 7.6 $¥%$ compared with the rule-based classifier, for coarse-grained and fine-grained categories, respectively. The results are shown in Figure 16.

7. CONCLUSION

Basic introduction on QA system has been introduced along with the example applications. Some approaches for question classification, including rule-based, machine-learning-based, and hybrid approaches, have been reported here. From the survey, the best one can achieve 95.0$% on coarse-grained category and 90.8% on fine-grained category.

The results of some of the mentioned classifiers are summarized in Figure 17. Note that, all of them used the two-layer taxonomy introduced in 3.2 and SVM classifiers are used. The results from shallow semantic structure and internet-based approach are not shown in Figure 17 because they are not comparable to others.

REFERENCES

[1] D. Jurafsky and J. H. Martin, *Speech and Language Processing*. Prentice Hall, 2009.

[2] X. Li and D. Roth, "Learning question classifiers," in *Proc. 19th Int. Conf. Computational linguistics - Volume 1*, ser. COLING'02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 1-7.

[3] E. Hovy *et al.*, "Question answering in webclopedia," in *Proc. 9th Text REtrieval Conf. (TREC-9)*, 2000, pp. 655-664.

[4] E. Hovy *et al.*, "A question/answer typology with surface text patterns," in *Proc. second Int. conf. Human Language Technology Research*, ser. HLT'02. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2002, pp. 247-251.

[5] J. Silva *et al.*, "From symbolic to sub-symbolic information in question classification,"

Artificial Intelligence Review, vol. 35, pp. 137-154, 2011, 10.1007/s10462-010-9188-4.

[6] M. Collins, “Head-driven statistical models for natural language parsing," *Comput. Linguist.*, vol. 29, no. 4, pp. 589-637, Dec. 2003.

[7] D. Zhang and W. S. Lee, “Question classification using support vector machines," in *Proc. 26th Annu. Int. ACM SIGIR Conf. Research and development in information retrieval*, ser. SIGIR'03. New York, NY, USA: ACM, 2003, pp. 26-32.

[8] M. Collins and N. Duffy, “Convolution kernels for natural language," in *Advances in Neural Information Processing Systems 14*. MIT Press, 2001, pp. 625-632.

[9] S. Bloehdorn and A. Moschitti, “Exploiting structure and semantics for expressive text kernels," in *Proc. Sixteenth ACM Conf. Information and Knowledge Management, CIKM 2007*, Lisbon, Portugal, November 2007, pp. 861-864.

[10] L. Marquez *et al.*, “A robust combination strategy for semantic role labeling," in *Proc. Conf. Human Language Technology and Empirical Methods in Natural Language Processing*, ser. HLT'05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 644-651.

[11] P. Kingsbury and M. Palmer, “From treebank to propbank," in *Proc. 3rd Int. Conf. Language Resources and Evaluation (LREC-2002)*, 2002.

[12] T. Solorio *et al.*, “A language independent method for question classification," in *Proc. 20th Int. Conf. Computational Linguistics*, ser. COLING'04. Stroudsburg, PA, USA: Association for Computational Linguistics, 2004.